Archive Table of Contents

알파고 이후 인공지능 판단 구조에 대한 해석 기록

– 정수 최적화 학습과 설명 체계 부재 가설

0. 기록 선언 (Record Declaration)

이 문서의 성격과 목적

본 문서는 인공지능 기술의 성능을 평가하거나, 특정 시스템의 우수성·위험성을 주장하기 위해 작성되지 않았다.

또한 미래에 대한 예측이나 정책·윤리적 권고를 제시하지 않는다.

이 기록의 목적은 단 하나이다.

인공지능이 '판단하는 것처럼 보이는 순간'에
실제로는 무엇이 발생하고 있는지를
판단 이전 단계의 구조로 되돌려 해석하고,
그 구조를 시간 좌표에 고정하는 것

즉, 본 문서는 **결론을 내리기 위한 문서가 아니라**,
판단이 성립하기 이전의 상태를 **기록하는 문서**다.

본 기록이 주장·예측·정책 제안이 아님을 명시

본 기록은 다음의 행위를 의도적으로 수행하지 않는다.

* 인공지능이 인간을 대체할 수 있는지에 대한 주장
* 인공지능을 규제하거나 허용해야 한다는 정책 제안
* 인공지능의 미래 위험이나 가능성에 대한 예측

이러한 논의들은 모두 **판단 이후의 층위**에 속한다.

그러나 본 기록이 다루는 대상은
아직 판단이 성립되기 이전의 단계,
즉 **판단이 '가능해 보이게 만드는 구조'** 자체이다.

PBT 관점에서 보면,

주장·예측·정책은 모두
이미 **판단이 내려진 상태**를 전제로 한다.
본 기록은 그 전제를 거부한다.

“설명 불가능성”을 해석 대상으로 삼는 이유

알파고 이후 인공지능에 대해 가장 자주 등장한 표현은
“블랙박스”와 “설명 불가능성”이다.
기존 논의는 주로 다음 중 하나에 머물렀다.
설명 방법이 아직 부족하다는 기술적 해석
설명 가능한 AI(XAI)를 추가하면 해결된다는 기대
설명할 수 없으므로 위험하다는 윤리적 경고
그러나 이 모든 접근은 공통된 전제를 가진다.
인공지능은 원래 판단하고 있으며,
다만 그 판단을 설명하지 못하고 있다.

본 기록은 이 전제를 보류한다.
PBT 관점에서 중요한 질문은 다음이다.
* 인공지능은 실제로 **판단을 하고 있는가**
* 아니면 판단처럼 보이는 **선택을 수행하고 있을 뿐인가**
* 설명 불가능성은 실패인가, 아니면 **구조적 결과인가**

본 기록은 “설명 불가능성”을
문제 해결의 대상이 아니라,
해석되어야 할 현상으로 다룬다.
즉,
설명이 없는 이유를 묻기 전에,
설명이 생성될 수 있는 구조가
존재했는지를 먼저 기록한다.

이 문서는 판단을 내리지 않는다.
판단을 유보하고,
판단 이전의 구조만을 고정한다.
그것이 본 기록의 역할이며,
PBT 언어로 표현된 이 문서의 위치이다.

1. 문제 인식: AlphaGo 이후에 남겨진 질문

알파고는 인간 최고 수준의 바둑 기사를 이기며,
인공지능이 분명히 **이전과 다른 단계**에 도달했음을 보여주었다.
이 성과는 단순한 승리가 아니라,
“기계가 어디까지 할 수 있는가”라는 질문을 현실로 끌어내렸다.
하지만 그 순간부터, 성과만큼이나 중요한 질문들이 함께 남았다.

AI는 왜 강해졌는가

알파고의 강함은 우연이 아니었다.
방대한 기보와 자기 대국을 통해, 바둑의 수많은 상황을
확률과 계산의 공간으로 흡수했기 때문이다.
이 과정에서 인공지능은
사람처럼 ‘생각하는 법’을 배운 것이 아니라,
어떤 선택이 이길 가능성을 높이는지를
끝없이 계산하고 축적했다.
그래서 알파고는
인간이 미처 보지 못한 수를 두기도 했고,
책에 없던 선택으로 승리를 쌓을 수 있었다.
강해진 이유는 분명하다.
계산과 확률에서 인간을 압도했기 때문이다.

왜 동시에 설명하지 못하는가

문제는 여기서 시작된다.
사람들은 자연스럽게 묻게 되었다.

* 왜 그 수를 두었는가?
* 무엇을 노린 선택이었는가?
* 그 판단의 기준은 무엇이었는가?

인간 기사라면
자신의 선택을 말로 풀어 설명할 수 있다.

하지만 알파고는 그렇지 않았다.
이때 많은 설명은
“아직 설명 기술이 부족하다”거나
“앞으로 설명 가능한 AI가 나오면 해결될 것”이라고 말한다.
그러나 이 기록은 다른 가능성을 본다.
알파고는 **설명을 못 한 것이 아니라**,
애초에 **설명을 만들 필요가 없는 방식으로 학습되었다**는 점이다.
즉, 알파고는
이유를 세워 판단하는 대신,
확률적으로 가장 유리한 선택을 계산해 실행했을 뿐이다.

기존 논의가 다루지 못한 공백

기존의 논의는 주로 두 방향에 머물렀다.
* AI가 얼마나 더 똑똑해질 것인가
* 설명하지 못하니 얼마나 위험한가
하지만 이 논의들 사이에는 중요한 공백이 있다.
AI가 실제로 ‘판단’을 하고 있는지,
아니면 판단처럼 보이는 ‘선택’을 수행하고 있는지에 대한 질문

이 구분이 없으면,
설명 불가능성은 단순한 결함처럼 보이고,
해결해야 할 문제로만 남는다.
본 기록은 이 지점에서 멈춰 서서 묻는다.
* 판단이 존재하지 않았다면, 설명은 왜 필요한가
* 설명이 없다면, 그것은 실패인가 구조의 결과인가
이 질문을 정리하지 않으면
알파고 이후의 인공지능 논의는
계속 같은 혼란을 반복하게 된다.
그래서 이 기록은
정답을 내리기보다,
그동안 건너뛰어졌던 질문의 자리를 먼저 고정하려 한다.

2. 바둑이라는 환경의 구조적 성질

알파고의 성과를 이해하기 위해서는,
먼저 바둑이라는 게임이 어떤 환경 위에 놓여 있는지를 살펴볼 필요가 있다.
알파고의 강함은 인공지능의 성능만으로 설명되지 않는다.
바둑이 가진 **구조적 성질**이 함께 작동했기 때문이다.
바둑은 대표적인 **완전 정보 게임**이다.
판 위의 모든 정보가 공개되어 있고,
우연 요소가 없으며,
규칙과 승리 조건이 변하지 않는다.
이러한 환경에서는
경험과 실력이 쌓일수록
선택이 무한히 늘어나기보다,
오히려 **비슷한 국면과 선택으로 수렴할 가능성**이 커진다.
처음에는 어디에나 둘 수 있을 것처럼 보이지만,
실력이 올라갈수록
명백히 불리한 수들이 하나씩 제거된다.
그 결과, 실제로 고려할 수 있는 선택지는
점점 좁아진다.
이 과정은 누군가가 규칙을 정해서 만든 것이 아니다.
같은 조건에서 같은 판단이 반복되면서,
확률적으로 살아남은 선택들만 남은 결과다.
우리가 흔히 말하는 '정석'이나 '패턴'은
이렇게 형성된 것이다.
실력이 낮을수록 선택은 많아 보인다.
이는 가능성이 넓어서가 아니라,
아직 피해야 할 선택을 모르기 때문이다.
반대로 실력이 높아질수록
둘 수 있는 수는 줄어들고,
한 수의 의미와 비용은 커진다.
알파고는 이 과정을
인간보다 훨씬 빠르게,
훨씬 넓은 범위에서 수행했다.
그래서 인공지능은 자유롭게 두는 것처럼 보였지만,
실제로는 **이미 좁아진 선택 공간을**

정확하게 계산해낸 것에 가깝다.
바둑은 종종
가장 자유도가 높은 게임으로 설명된다.
판이 넓고 규칙이 단순하며,
창의적인 수가 가능해 보이기 때문이다.
그러나 이 자유는
구조 바깥의 자유가 아니라,
구조 안에서 허용된 자유다.
규칙은 고정되어 있고,
승리 조건은 명확하며,
잘못된 선택은 즉시 비용으로 돌아온다.
이러한 환경에서는
아무리 창의적으로 보이는 수라도
시간이 지나면
확률적으로 검증되고 평가된다.
알파고는 이 검증 과정을
인간보다 훨씬 더 많이,
훨씬 더 빠르게 수행했을 뿐이다.
따라서 알파고의 승리는
자유로운 사고가 인간을 넘어섰다는 증거라기보다,
구조적으로 닫힌 환경에서
확률 수렴을 극대화한 결과로 이해할 수 있다.
이 점을 이해하지 못하면,
알파고의 성과를
인공지능이 ‘판단’을 획득한 사건으로
오해하게 된다.
이 문서는 그 오해를 피하기 위해,
바둑이라는 환경의 구조부터
차분히 되짚는다.

3. 인간 기보와 타계 연구의 역할

알파고의 학습에서 자주 오해되는 지점은
인공지능이 인간의 기보를 **정답처럼 외웠다**는 인식이다.
그러나 인간 기보와 타계 연구는
정답 집합이 아니라 **출발점을 정렬해주는 기준선**에 가깝다.

기보는 정답 집합이 아닌 초기 좌표계

인간의 기보는
“이 상황에서는 반드시 이 수가 맞다”를 모아놓은 목록이 아니다.
실제로 기보에는 실수도 있고,
시대에 따라 평가가 바뀐 수들도 섞여 있다.
그럼에도 기보가 중요한 이유는 분명하다.
기보는 바둑판 위에서
인간이 무엇을 중요하게 보고,
어디를 먼저 살펴보는지를 보여준다.
즉, 기보의 역할은
정답을 제공하는 것이 아니라,
판을 해석하는 **초기 좌표계를 형성**하는 데 있다.
알파고는 이 좌표계를 통해
무작위로 탐색하지 않고,
인간이 이미 걸러낸 방향에서
계산을 시작할 수 있었다.

타계책과 연구 수의 구조적 한계

타계책과 연구 수는
인간이 오랜 시간 동안 축적한 사고의 결과다.
특정 국면에서 어떤 교환이 유리한지,
어떤 형태가 반복적으로 등장하는지를 정리해 놓은 것이다.
하지만 이 연구들은 본질적으로

언어와 개념에 의존한다.

* 두터움
* 세력
* 안정
* 주도권

이런 개념들은
인간에게는 유용하지만,
정확한 수치로 고정되지는 않는다.
그래서 타게 연구는
많은 경우 “대체로 그렇다”는 표현에 머문다.
이것이 인간 사고의 강점이자 동시에 한계다.
알파고는 이 개념들을 이해하지 않는다.
대신, 타게 연구가 다루지 못한 수많은 경우까지 포함해
결과로만 검증한다.
이 지점에서
인간 연구의 한계가 드러난다.
사람은 설명 가능한 범위까지만 정리하지만,
인공지능은 설명되지 않는 영역까지 계산한다.

인간 사고의 요약물로서의 기보

이 관점에서 보면,
기보는 인간 사고의 기록이 아니라
인간 사고의 요약본에 가깝다.

* 무엇을 보고
* 무엇을 중요하게 여겼고
* 무엇을 피하려 했는지

이 모든 것이
수 하나하나에 압축되어 남아 있다.
알파고는 이 요약물을
의미로 읽지 않는다.

개념으로 해석하지도 않는다.
기보는 알파고에게
“생각의 기록”이 아니라,
확률 공간을 형성하는 데이터로 작동한다.
그 결과,
알파고는 인간이 남긴 사고의 흔적을 바탕으로
인간과는 전혀 다른 방식의 계산을 수행할 수 있었다.
이 장에서 중요한 점은 하나다.
인간 기보는 인공지능에게
답을 가르친 것이 아니라,
어디서부터 계산을 시작할지를 알려주었을 뿐이라는 점이다.

이 이해가 있어야만,
알파고의 수가
‘암기된 인간의 수’가 아니라
재연산된 선택이라는 사실이 분명해진다.

4. 알파고의 학습 구조에 대한 해석

알파고의 강함은 ‘생각을 잘해서’가 아니라,
판을 다루는 방식 자체가 인간과 달랐기 때문이다.
이 장에서는 알파고가 바둑을 어떻게 받아들이고,
어떤 기준으로 수를 선택했는지를 구조적으로 해석한다.

국면의 수치화와 벡터화

인간은 바둑판을 보며
세력, 두터움, 약점 같은 **의미**를 먼저 떠올린다.
그러나 AlphaGo 는
바둑판을 그런 방식으로 보지 않는다.
알파고에게 한 판의 국면은

의미 있는 장면이 아니라
수많은 숫자로 이루어진 상태값이다.

* 돌의 위치
* 연결 관계
* 빈 칸의 분포

이 모든 요소는
하나의 그림이나 개념이 아니라
벡터와 수치 배열로 변환된다.
이 과정에서 중요한 점은,
무엇이 '중요한지'를 사람이 지정하지 않는다는 것이다.
국면은 해석되지 않고,
그대로 **계산 가능한 상태**로만 입력된다.

승률 예측과 기대값 최대화

이렇게 수치화된 국면 위에서
알파고가 하는 일은 단순하다.

* 이 상태에서
* 이 수를 두면
* **이길 확률이 얼마나 되는가**

알파고는
"이 수가 공격적인가",
"이 수가 두터운가"를 묻지 않는다.
오직
승률이 얼마나 변하는지만을 계산한다.
여러 수를 동시에 비교한 뒤,
가장 기대값이 높은 선택을 고른다.
이 과정은 반복적이고 일관되다.
중요한 점은 여기서
'판단'이 개입하지 않는다는 것이다.

* 옳고 그름의 기준도 없고
* 의미에 대한 해석도 없으며
* 선택을 설명해야 할 필요도 없다

그저 **확률적으로 가장 유리한 값**이 선택된다.

자기대국 강화학습의 의미

알파고의 학습에서 결정적인 단계는
자기 자신과의 대국, 즉 **자기대국 강화학습**이다.
이 단계에서 알파고는
인간 기보를 더 이상 참고하지 않는다.
오직 다음의 기준만 남는다.

* 이 선택이
* 결국 이겼는가, 졌는가

이 결과를 바탕으로
각 국면의 가치가 조금씩 수정된다.
이 과정이 수백만 번 반복되며,
확률 분포는 점점 정교해진다.
여기서 중요한 점은,
알파고가 **이유를 배우지 않는다는 사실**이다.

* 왜 이겼는지
* 어떤 의미에서 유리했는지

이런 질문은 학습 과정에 포함되지 않는다.
오직 결과만이 남고,
그 결과가 다시 계산 구조를 조정한다.
즉, 자기대국은
알파고에게 통찰을 준 것이 아니라,
확률 공간을 더욱 촘촘하게 만들었을 뿐이다.

이 장에서 확인할 수 있는 핵심은 분명하다.
알파고의 학습은

이해를 깊게 한 과정이 아니라,
정수화된 국면 위에서
기대값 계산을 극도로 정밀하게 만든 과정이다.

이 구조를 이해하지 못하면,
알파고의 선택을
'판단'이나 '사고'로 오해하게 된다.

5. 정수 최적화 학습 가설

앞선 장들에서 살펴본 학습 과정을 종합하면,
알파고의 강함은 하나의 가설로 정리될 수 있다.
본 기록은 이를 **정수 최적화 학습 가설**이라 부른다.
이 가설의 핵심은 단순하다.
알파고는 의미를 이해해 판단하는 구조를 갖지 않았고,
대신 **정수(확률) 공간에서 선택을 최적화하는 구조**를 고정했다는 점이다.

"국면 → 의미 → 판단" 구조의 부재

인간의 바둑 사고는 대체로 다음의 흐름을 따른다.
* 국면을 본다
* 그 국면을 의미로 해석한다
* 해석된 의미를 바탕으로 판단한다
여기에는 항상 **개념의 층위**가 개입한다.
세력, 두터움, 안정, 약점과 같은 개념들이
판단의 중간 다리 역할을 한다.
그러나 AlphaGo 의 학습 구조에는
이 중간 단계가 존재하지 않는다.
알파고는 국면을 보지만,
그 국면을 '의미 있는 상태'로 해석하지 않는다.
따라서 판단에 이르기 위한

개념적 경로 자체가 형성되지 않는다.
이것은 기능의 부족이 아니라,
구조 설계의 선택이다.

"국면 → 정수 → 선택" 구조의 고정

알파고의 실제 흐름은 다음과 같이 요약된다.
* 국면을 입력한다
* 국면을 수치와 벡터로 변환한다
* 각 선택의 승률을 계산한다
* 가장 기대값이 높은 선택을 실행한다

이 구조에서 중요한 점은,
판단이라는 단계가 개입할 여지가 없다는 것이다.

* 옳고 그름을 따지지 않고
* 의미를 비교하지 않으며
* 설명을 생성하지 않는다

선택은 언제나
정수 값의 크기 비교로 결정된다.
이 흐름이 반복되면서,
알파고의 내부에는
"이 국면에서는 이 선택이 유리하다"는
정수적 대응 관계만이 축적된다.

정수(확률) 공간에서의 재연산 개념

이 가설에서 말하는 '새로운 수'는
창의적 발상이나 통찰의 결과가 아니다.
알파고의 수는
인간 기보에 기록된 타계 방법을
그대로 재현한 것도 아니다.

그 과정은 다음과 같다.

* 인간의 기보와 연구 수는
하나의 확률 지형으로 변환된다

* 각 국면은
그 지형 위의 좌표로 매핑된다

* 선택은
그 좌표에서 다시 계산된 결과로 도출된다

즉, 알파고의 수는
기억의 산물이 아니라,
정수 공간에서의 재연산 결과다.
이 때문에 인간이 보기에
낯설고 설명되지 않는 수가 등장하지만,
알파고 내부에서는
일관된 계산의 연속일 뿐이다.

이 장에서 제시한 정수 최적화 학습 가설은
알파고의 강함과 동시에
설명 불가능성이 발생한 이유를
하나의 구조로 묶는다.

알파고는 판단을 학습하지 않았다.
국면을 정수화했고,
정수의 최대값을 선택하는 법을 학습했을 뿐이다.

6. 패턴의 재현이 아닌 재연산으로서의 '새로운 수'

알파고의 대국에서 가장 큰 충격을 준 장면들은
기존 기보나 타게 연구에서 거의 등장하지 않았던 수들이었다.
이 수들은 흔히 "창의적"이거나 "신의 한 수"로 불렸지만,
그 성격을 정확히 이해하려면 다른 관점이 필요하다.

인간이 두지 않은 수가 등장하는 이유

알파고가 인간이 두지 않았던 수를 둘 수 있었던 이유는,
인간의 수를 **그대로 따라 했기 때문이 아니라**,
인간의 수를 **확률 구조로 바꾼 뒤 다시 계산했기 때문**이다.
인간의 기보와 타게 연구는
특정 국면에서 자주 선택된 수들을 중심으로 정리되어 있다.
이는 인간 사고의 효율적인 요약이지만,
동시에 인간이 **잘 보지 않았던 선택지**를 자연스럽게 배제한다.
알파고는 이 배제를 그대로 따르지 않는다.
인간이 거의 두지 않았던 수라도,
승률 계산에서 의미 있는 값을 가진다면
동등한 후보로 다시 평가된다.
그 결과,
인간의 경험과 관습에서는 밀려났던 선택이
확률 계산을 통해 다시 살아난다.
이 지점에서 "새로운 수"가 등장한다.

새로움과 이해의 분리

알파고의 수가 새로워 보이는 이유는 분명하다.
인간의 개념 체계 안에서
그 수를 설명할 언어가 없기 때문이다.
그러나 이 새로움은
이해의 결과가 아니다.

* 알파고는 그 수의 의미를 알지 못하고
* 왜 낯설게 보이는지도 알지 못하며
* 기존 개념을 깨겠다는 의도도 없다

새로움은 단지
정수(확률) 공간에서의 계산 결과로 나타난다.
즉, 알파고의 수는
“이해해서 새로워진 수”가 아니라,
이해를 거치지 않고도 도달한 수다.
이 지점에서
인간이 느끼는 감탄과
알파고 내부의 계산 사이에는
본질적인 간극이 존재한다.

창작이 아닌 계산으로서의 혁신

알파고의 수를
인간의 창작과 동일한 범주로 놓는 것은
오해에 가깝다.
창작에는 보통 다음이 포함된다.

* 기존 개념의 재구성
* 의미의 전환
* 의도나 문제의식

하지만 알파고의 선택에는
이 어떤 요소도 포함되지 않는다.
알파고가 만들어낸 변화는
발상의 전환이 아니라,
계산 범위의 확장이다.
인간이 다루지 못했던 수많은 경우를
동시에 계산하고,
그중 결과가 좋은 선택을
조용히 끌어올렸을 뿐이다.
이 점에서 알파고의 혁신은

창작의 혁신이 아니라,
계산 방식의 혁신으로 이해하는 것이 더 정확하다.

이 장의 핵심은 다음 한 문장으로 정리된다.
알파고의 새로운 수는
패턴을 깨서 나온 것이 아니라,
패턴을 확률로 환원한 뒤
다시 계산한 결과로 등장했다.

이 이해가 있어야만,
알파고의 새로움이
'판단의 증거'가 아니라
정수 최적화 구조의 부산물임을 분명히 볼 수 있다.

7. 설명 체계가 생성되지 않는 구조적 이유

알파고 이후 인공지능을 둘러싼 논의에서
가장 자주 등장한 요구는 "설명을 붙이자"는 것이었다.
그러나 본 장의 결론은 분명하다.
설명은 추가로 얹을 수 있는 기능이 아니라,
처음부터 구조 안에서 생성되어야 하는 층위다.

설명은 왜 '추가'될 수 없는가

설명은 결과를 말로 풀어내는 행위가 아니다.
설명은 다음의 과정을 전제로 한다.

* 선택 이전에 개념이 형성되고
* 그 개념들 사이의 관계가 정리되며
* 선택이 그 관계 위에서 정당화되는 과정

즉, 설명은 **판단의 부산물**이다.
알파고의 학습 구조에는
이 전제가 성립하지 않는다.
선택은 언제나 계산의 결과로 바로 도출되며,
그 이전에 “왜 이 선택이 의미 있는가”를 묻는 단계가 없다.
이 상태에서 설명을 붙이려 하면,
실제로는 다음 중 하나가 된다.

* 결과를 인간의 언어로 **사후 번역**하거나
* 일부 수치를 시각적으로 **꾸며 보여주는 것**

이것은 설명처럼 보일 수는 있지만,
선택을 낳은 내부 구조를 설명하지는 못한다.
그래서 설명은
나중에 덧붙일 수 없는 것이다.

개념 계층 부재의 의미

인간의 사고에는
항상 개념의 계층이 존재한다.

* 이 수는 공격적인가
* 이 선택은 안전한가
* 이 교환은 장기적으로 어떤 의미를 가지는가

이 개념들은
선택을 가능하게 하는 동시에,
선택을 설명 가능하게 만든다.
그러나 알파고의 구조에는
이 개념 계층이 존재하지 않는다.
국면은 수치로 변환되고,
수치는 바로 비교된다.
그 사이에 머무를 **의미의 층**이 없다.
이것은 단순한 결핍이 아니다.
개념 계층을 만들지 않았기 때문에,

알파고는 훨씬 빠르고 안정적으로
정수 최적화를 수행할 수 있었다.
즉,
설명 불가능성은 실패가 아니라 선택의 결과다.

블랙박스의 발생 조건

이 지점에서
'블랙박스'라는 말의 의미를 다시 정리할 필요가 있다.
블랙박스는
정보가 숨겨져 있어서 생기는 것이 아니다.
또는 기술이 미숙해서 생기는 것도 아니다.
블랙박스는 다음 조건이 동시에 충족될 때 발생한다.
* 국면이 의미가 아닌 정수로 표현되고
* 선택이 판단이 아닌 계산으로 이루어지며
* 개념 계층이 생성되지 않는 구조가 고정될 때
이때 외부에서 보이는 것은
입력과 출력뿐이다.
그 사이에는
설명될 수 있는 판단 경로가 존재하지 않는다.
따라서 블랙박스는
우연히 생긴 문제가 아니라,
정수 최적화 학습 구조가 만들어내는 필연적 결과다.

이 장에서 정리된 결론은 단순하다.
알파고가 설명하지 못하는 이유는
설명을 감추고 있기 때문이 아니라,
설명이 생성될 수 없는 구조로 학습되었기 때문이다.

이 인식을 받아들이지 않으면,
설명 불가능성은 계속해서
"언젠가 해결될 기술적 문제"로 오해된다.

8. 블랙박스 해석의 재정의

알파고 이후 인공지능을 설명할 때,
“블랙박스”라는 표현은 거의 자동처럼 사용되어 왔다.
그러나 이 표현은 종종 **문제를 흐리게 만든다**.
본 장에서는 블랙박스를
“알 수 없는 상태”가 아니라
“그럴 수밖에 없었던 구조”로 재정의한다.

“알 수 없음”이 아닌 “그럴 수밖에 없음”

일반적으로 블랙박스는
“내부를 아직 이해하지 못했다”는 뜻으로 쓰인다.
즉, 시간이 지나고 기술이 발전하면
언젠가는 열릴 상자처럼 취급된다.
하지만 AlphaGo 의 경우,
이 해석은 정확하지 않다.
알파고의 내부가 보이지 않는 이유는
복잡해서가 아니라,
보여줄 판단 경로 자체가 없기 때문이다.

* 판단을 거치지 않고
* 개념을 만들지 않으며
* 의미를 정리하지 않는 구조에서

외부로 드러날 수 있는 것은
입력과 출력뿐이다.
따라서 이 블랙박스는
“아직 알 수 없음”이 아니라,
“구조적으로 알 수 없게 설계된 상태”다.

기술 부족설 vs 구조 필연설

기존 논의의 많은 부분은
설명 불가능성을 다음과 같이 해석해 왔다.

* 모델이 아직 충분히 발전하지 않았다
* 설명 모듈이 붙지 않았다
* 해석 기술이 뒤따라오지 못했다

이것이 **기술 부족설**이다.
반면 본 기록이 제시하는 관점은 다르다.
설명 불가능성은 기술의 미성숙이 아니라,
정수 최적화 학습 구조가 만들어낸 필연적 결과라는 것이다.

* 학습은 수치 공간에서 이루어지고
* 선택은 기대값 비교로 결정되며
* 그 사이에 설명을 생성할 층위가 없다

이 구조가 유지되는 한,
아무리 많은 해석 기법을 덧붙여도
설명은 본질적으로 사후적일 수밖에 없다.
즉, 문제는
“설명을 못 하는 AI”가 아니라,
“설명을 만들지 않는 AI”다.

설명 불가능성의 위치 고정

이 지점에서 중요한 작업은
설명 불가능성을 해결 과제로 두는 것이 아니라,
그 위치를 정확히 고정하는 것이다.
설명 불가능성은:

* 오류도 아니고
* 결함도 아니며
* 우연히 생긴 문제도 아니다

그것은 특정한 학습 구조가 작동할 때
반드시 따라오는 **구조적 특성**이다.
이 위치를 고정하지 않으면,
논의는 반복된다.

* "언젠가는 설명될 것이다"라는 기대
* "설명되지 않으니 위험하다"는 불안
* "설명 기능을 추가하자"는 처방

그러나 구조를 바꾸지 않는 한,
결과는 달라지지 않는다.
본 기록은 여기서 멈춰 서서 선언한다.
알파고의 블랙박스는
아직 열리지 않은 상자가 아니라,
열릴 수 없도록 만들어진 구조적 결과다.

이 재정의가 있어야만,
설명 가능성·판단·책임에 대한 논의가
현실적인 출발점을 가질 수 있다.

9. 가설의 성격과 반증 가능성

앞선 장들에서 제시한 해석은
알파고의 성과를 설명하기 위한 **하나의 관점**이지,
최종 결론이나 규범적 판단이 아니다.
이 장에서는 이 가설이 어떤 위치에 있으며,
어떤 조건에서 유지되거나 수정될 수 있는지를 명확히 한다.

해석 가설로서의 위치

본 기록에서 제시한 정수 최적화 학습 가설은
새로운 기술 사실을 주장하지 않는다.

알파고의 학습 방식, 강화학습 구조, 승률 계산 등은
이미 공개된 자료와 연구를 기반으로 한다.
이 가설의 역할은
사실을 추가하는 것이 아니라,
이미 알려진 사실들을 **하나의 구조로 묶어 해석**하는 데 있다.
즉, 이 가설은 다음과 같은 성격을 가진다.
* 기술적 구현에 대한 주장 ✕
* 윤리적·정책적 결론 ✕
* 미래 예측 ✕
대신,
왜 강해졌는지와
왜 설명하지 못했는지를
동시에 설명하려는 해석적 틀

이라는 위치에 놓인다.
따라서 이 가설은
맞고 틀림의 문제가 아니라,
얼마나 잘 설명하는가의 문제로 평가되어야 한다.

반증 조건 명시

본 가설은 반증 가능하다.
이는 기록으로서 중요한 조건이다.
다음과 같은 변화가 실제로 관찰된다면,
본 가설은 수정되거나 폐기되어야 한다.

* 인공지능이 스스로 개념을 생성하고, 그 개념을 유지·수정하며

* 선택 이전에 이유와 기준을 내부적으로 구성하고

* 그 기준에 따라 선택을 보류하거나 변경하며

* 선택 이후에 결과에 대한 책임 경로를 구조적으로 연결하는 경우

이때 설명은
사후적 해설이 아니라,
선택을 만들어낸 **내부 과정의 일부**가 된다.
이러한 구조가 실현된다면,
“설명 체계가 생성되지 않는다”는 본 가설은
더 이상 유지될 수 없다.
반대로,
설명 기능이 외부 모듈이나 시각화로만 제공되고
선택 구조 자체가 변하지 않는 한,
본 가설은 여전히 유효하다.

향후 기술 발전과의 관계

본 가설은
인공지능 기술의 발전 가능성을 부정하지 않는다.
다만, 발전의 방향을 구분한다.

* 계산 능력의 증가
* 데이터 규모의 확대
* 탐색 효율의 개선

이러한 발전은
정수 최적화 구조를 더욱 강하게 만들 수는 있어도,
설명 체계를 **자연스럽게 생성하지는 않는다**.
설명 체계의 생성은
성능 향상의 연장선이 아니라,
구조의 전환을 요구한다.
따라서 향후 기술 발전은
두 가지 방향 중 하나로 나뉠 가능성이 크다.

* 정수 최적화 구조를 극대화하는 방향
* 판단·개념·설명 층위를 포함하는 새로운 구조로의 전환

본 가설은

어느 쪽이 옳다고 주장하지 않는다.
다만, 두 방향을 **구조적으로 구분**하고,
현재의 알파고 계열 시스템이
어느 위치에 있는지를 기록한다.

이 장의 요지는 다음과 같다.
본 가설은 주장이나 예언이 아니라,
특정 시점의 구조를 해석적으로 고정한 기록이며,
반증 가능성을 전제로 열린 상태로 유지된다.

10. 기록의 범위와 한계

이 장은 본 기록이 **무엇을 말하는 문서인지**보다,
무엇을 말하지 않기로 선택한 문서인지를 분명히 하기 위해 존재한다.
이 경계선이 명확할수록, 기록은 오독되지 않는다.

이 기록이 다루지 않는 것들

본 기록은 다음의 주제를 **의도적으로 다루지 않는다**.
* 인공지능의 윤리적 옳고 그름
* 인공지능의 위험성에 대한 경고나 낙관
* 인공지능 규제 또는 활용에 대한 정책 제안
* 인공지능이 인간을 대체할 가능성에 대한 판단
이러한 주제들은 모두
판단이 이미 내려진 이후의 층위에 속한다.
본 기록은 그 이전,
즉 **판단이 가능해 보이게 만드는 구조**까지만을 다룬다.
판단 이후의 결론을 이 문서에 요구하는 것은
이 기록의 목적과 범위를 벗어난다.

윤리·정책·책임 논의와의 거리

윤리, 정책, 책임 논의는 중요하다.
그러나 이 논의들은 공통적으로
다음 전제를 필요로 한다.

* 누가 판단했는가
* 무엇을 근거로 판단했는가
* 그 판단에 책임을 물을 수 있는가

본 기록이 다루는 대상은
이 전제가 **성립하기 이전의 구조**다.
즉, 이 문서는
윤리나 책임 논의에 **반대**하지 않으며,
그 논의를 **회피**하지도 않는다.
다만,
판단 구조가 어떻게 만들어졌는지에 대한
기본 해석 없이 진행되는 논의는
불필요한 혼란을 낳을 수 있음을 지적할 뿐이다.
이 기록은
윤리·정책 논의로 넘어가기 전에
반드시 한 번 거쳐야 할 해석의 지점을
고정하는 역할을 맡는다.

오독 방지를 위한 경계선

본 기록은 다음과 같은 방식으로
오독되어서는 안 된다.

* “AI는 설명할 수 없으니 위험하다”라는 주장
* “설명 없는 AI는 사용해서는 안 된다”는 결론
* “알파고는 생각하지 못한다”는 가치 판단

이 기록이 말하는 것은
찬반이나 평가가 아니다.

설명 불가능성은 결함인지,
아니면 특정 구조에서 필연적으로 나타나는 특성인지를
구분하기 위한 해석적 시도다.

따라서 본 문서에서의
'설명 불가능성'은
비난의 대상도, 옹호의 대상도 아니다.
그것은
구조가 만들어낸 결과로서
그 위치만 정확히 고정된다.

이 장의 요지는 간단하다.
이 기록은
판단을 내리는 문서가 아니라,
판단이 어디에서 가능해지고
어디에서 멈추는지를 구분하기 위한 문서다.

이 경계선을 넘는 해석은
독자의 몫이지만,
그 경계선 자체는
이 기록 안에서 분명히 남겨진다.

11. 결론: 설명 없는 최적화가 남긴 것
이 기록은 알파고의 성과를 찬양하거나 비판하기 위해 쓰이지 않았다.
대신, 알파고가 무엇을 **보여주었고**, 무엇을 **보여주지 않았는지**를
구조적으로 구분하는 데 목적이 있었다.
이 결론은 그 구분을 최종적으로 고정한다.

알파고가 증명한 것과 증명하지 않은 것

AlphaGo 는 분명히 중요한 사실을 증명했다.
* 계산과 확률 최적화만으로도
인간 최고 수준의 성과에 도달할 수 있다는 점

* 완전 정보 환경에서는
설명 없이도 성능이 극대화될 수 있다는 점

이것은 기술적으로도, 역사적으로도 명확한 성과다.
그러나 동시에 알파고는
다음의 것들을 **증명하지 않았다**.

* 인공지능이 판단을 수행할 수 있다는 점
* 인공지능이 의미를 이해한다는 점
* 인공지능이 선택에 대한 이유를 구성할 수 있다는 점

알파고의 승리는
'이해의 승리'가 아니라,
설명 없는 최적화가 어디까지 가능한지를 보여준 사례였다.
이 두 영역을 구분하지 않으면,
성과는 과대평가되거나 잘못 해석된다.

판단·설명·책임 논의로 이어지는 구조적 단서

본 기록이 반복해서 강조한 핵심은 하나다.
설명이 없다는 사실보다 중요한 것은,
설명이 생성될 구조가 존재했는가이다.

알파고의 구조에는
* 판단을 형성하는 개념 계층이 없었고
* 판단을 보류하거나 수정하는 단계가 없었으며
* 선택 이후 책임을 연결할 경로도 없었다

따라서 설명의 부재는
윤리적 문제 이전에
구조적 결과로 이해되어야 한다.
이 인식은
향후 판단·설명·책임을 요구하는 영역에서
인공지능을 어떻게 다루어야 하는지에 대해
중요한 단서를 제공한다.

* 어디까지는 설명 없는 최적화가 허용되는가
* 어디부터는 판단 구조가 필요해지는가
* 책임을 묻기 위해 무엇이 선행되어야 하는가

이 질문들은
알파고를 넘어서,
현실 세계로 이어지는 지점에서 비로소 의미를 갖는다.

본 기록의 참조적 가치

이 문서는 해답을 제시하지 않는다.
또한 독자에게 어떤 결론을 강요하지도 않는다.
본 기록의 가치는
다음 한 지점에 있다.

설명 없는 최적화와
판단이 요구되는 영역을
구조적으로 구분할 수 있는 기준점을 남겼다는 것

시간이 지나
기술이 더 발전하고,
설명 가능한 AI나 새로운 구조가 등장하더라도,
이 기록은 하나의 참조점으로 남을 수 있다.

* 어떤 구조에서 설명이 생성되지 않았는지
* 그것이 실패가 아니라 선택의 결과였는지
* 판단 논의는 어디서부터 시작되어야 하는지

이 질문들이 다시 등장할 때,
본 기록은
“그 시점에 이렇게 해석했다”는
구조적 기록으로 기능할 것이다.

이로써 이 문서는 끝난다.
이 기록은 예측이 아니라 고정이다.
설득이 아니라 분리다.
결론이 아니라,
판단 이전의 구조를 남긴다.

그 역할은 여기까지다.

Appendix A. 핵심 명제 요약

본 부록은 본문 전체를 관통하는 핵심 논지를
명제 형태로 간결하게 고정하기 위해 작성되었다.
아래의 명제들은 주장이나 결론이 아니라,
본 기록이 채택한 **해석의 기준점**이다.

A-1. 정수 최적화 학습 정의

정수 최적화 학습이란,
국면을 의미나 개념으로 해석하지 않고
수치(확률, 기대값)로 변환한 뒤,
그 값의 크기 비교를 통해 선택을 수행하는 학습 구조를 말한다.

이 구조에서 학습되는 것은 다음이다.
* 국면과 선택 간의 정수적 대응 관계
* 선택이 가져온 결과에 따른 확률 조정

반면, 다음은 학습되지 않는다.
* 선택의 의미
* 판단의 이유
* 개념적 기준

정수 최적화 학습은
이해를 전제로 하지 않으며,
설명 가능성을 목표로 하지 않는다.

A-2. 설명 체계 부재 명제

설명 체계는 선택 이후에 덧붙일 수 있는 기능이 아니라,
판단 과정 내부에서만 생성될 수 있는 구조적 층위다.

따라서서 다음이 성립한다.

* 판단을 구성하는 개념 계층이 없고
* 선택이 계산 결과로 즉시 도출되며
* 이유를 구성·유지·검토하는 단계가 존재하지 않는 경우

설명은 생성되지 않는다.
이때 발생하는 설명 불가능성은
기술적 결함이나 정보 부족이 아니라,
학습 구조가 만들어낸 필연적 결과다.

A-3. 패턴 수렴 게임 명제

완전 정보 게임에서는
실력과 탐색 밀도가 증가할수록
선택은 무작위로 확산되지 않고
확률적으로 수렴한다.

바둑과 같은 환경에서는:

* 규칙과 승리 조건이 고정되어 있고
* 우연 요소가 없으며
* 반복된 검증을 통해 불리한 선택이 제거된다

그 결과,
특정 국면에서는 유사한 선택들이
높은 확률로 반복된다.
이 수렴은
자유의 상실이 아니라,
구조가 허용한 범위 안에서의 안정화다.
정수 최적화 학습은
이 수렴 과정을 인간보다 빠르고 넓게 수행할 수 있으며,
그 결과 성능은 극대화되지만
설명 체계는 동반되지 않는다.

Appendix A 요약 문장

정수 최적화 학습은
패턴 수렴 환경에서 강력한 성능을 보이지만,
그 구조상 설명 체계를 생성하지 않는다.

이 명제들은
본 기록 전체를 관통하는 **해석의 좌표**로서,
이후 논의에서 반복적으로 참조될 수 있다.

Appendix B. 용어 정의

본 부록은 본 기록에서 사용된 핵심 용어들이
일상적 의미나 기술적 관용과 혼동되지 않도록
이 문서 내부에서의 사용 의미를 고정하기 위해 작성되었다.

B-1. 국면

국면이란,
특정 시점의 바둑판 상태 전체를 의미한다.
여기에는 다음이 모두 포함된다.
* 돌의 배치
* 연결 관계
* 빈 칸의 분포
* 수순에 의해 형성된 현재 상태
본 기록에서 국면은
“의미 있는 장면”이나 “전략적 상황”이 아니라,
선택의 입력으로 주어지는 상태값을 가리킨다.
즉, 국면은 해석의 대상이기 이전에
연산의 출발점이다.

B-2. 정수 최적화

정수 최적화란,
상태나 선택을 의미·개념으로 판단하지 않고,
수치화된 값(확률, 기대값 등)의 크기를 비교하여
가장 큰 값을 선택하는 방식을 말한다.
이 구조의 특징은 다음과 같다.

* 의미 판단을 거치지 않는다
* 옳고 그름을 구분하지 않는다
* 설명을 생성하지 않는다

정수 최적화는
“왜”가 아니라
“얼마나 유리한가”에만 반응한다.
본 기록에서 정수 최적화는
알파고 학습 구조의 핵심 작동 원리로 사용된다.

B-3. 설명 체계

설명 체계란,
선택 이전에 형성된 개념과 기준을 바탕으로
선택의 이유를 구성·유지·검토할 수 있는 구조를 말한다.
설명 체계에는 다음이 포함된다.

* 개념의 생성과 유지
* 개념 간 관계의 정리
* 선택을 정당화하는 언어적·논리적 경로

중요한 점은,
설명 체계는 **선택 이후에 덧붙이는 기능이 아니라**,
판단 과정 내부에서만 생성될 수 있다는 것이다.
본 기록에서 말하는
“설명 불가능성”은
설명 체계의 부재를 의미한다.

B-4. 판단 vs 선택

본 기록에서는
판단과 **선택**을 명확히 구분한다.

판단

* 개념을 전제로 한다
* 기준과 이유를 구성한다
* 보류·수정·책임이 가능하다
* 설명 체계를 포함한다

판단은
의미를 다루는 행위다.

선택

* 수치 비교를 전제로 한다
* 결과만을 기준으로 한다
* 보류나 책임 개념이 없다
* 설명을 요구하지 않는다

선택은
연산의 결과다.
본 기록에서 알파고가 수행한 것은
판단이 아니라 **선택**이며,
이 구분이 설명 불가능성의 구조적 원인을 이해하는 핵심이다.

Appendix B 요약 문장

본 기록에서 사용되는 용어들은
일상적 의미가 아니라,
구조를 구분하기 위한 기준어로 정의된다.

이 정의는
본문 전체를 해석하는 공통 좌표로 사용된다.

Appendix C. Status Note

본 부록은 이 기록의 **법적·학문적·해석적 위치**를 명확히 하기 위해 작성되었다.
아래의 표기는 이 문서가 어떤 효력을 **의도하지 않는지**를 함께 고정한다.

Interpretive Hypothesis

본 기록은 해석 가설(Interpretive Hypothesis)이다.
새로운 기술 사실이나 실험 결과를 제시하지 않으며,
이미 공개된 자료와 관측된 현상을
하나의 구조적 관점으로 **재해석**한다.
이 가설은 다음을 전제로 한다.

* 특정 시점에서 관측된 구조에 대한 해석임
* 반증 가능성을 열어 둠
* 향후 기술 변화에 따라 수정·보완·폐기될 수 있음

따라서 본 기록은
정답을 선언하지 않고,
해석의 기준점을 제시하는 데 그 목적이 있다.

Non-binding

본 기록은 **비구속적(Non-binding)** 문서다.
어떠한 의무, 규범, 기준도 강제하지 않는다.
특히 다음을 포함하지 않는다.

* 법적 책임 판단
* 정책적 권고
* 윤리적 평가
* 기술 채택 또는 배제에 대한 지침

본 기록의 내용은
판단을 유도하기 위한 근거가 아니라,
판단 이전 구조를 이해하기 위한 참고로만 사용된다.

Reference Only

본 기록은 참조용(Reference Only)으로 남겨진다.
설득이나 합의를 목표로 하지 않으며,
논쟁을 촉발하기 위한 문서도 아니다.
이 문서의 역할은 다음과 같다.
* 특정 시점에 존재했던 해석을 고정
* 이후 등장하는 논의들과의 비교 기준 제공
* 구조 변화 여부를 판단하기 위한 참조점 역할
즉, 이 기록은
미래의 결론을 대신하지 않고,
미래의 논의가 돌아와 확인할 좌표로 존재한다.

Status Summary

본 기록은
해석 가설이며,
비구속적이고,
참조용으로만 남겨진다.

이 세 가지 표기는
본 문서 전체의 성격을 규정하는
최종 상태 선언이다.